Comparative Study and Predictive Modeling of Machine Learning Algorithms for Traffic Flow in Smart Transportation Systems

S. Santhi Rupa [1*], M. Jyothi Sai Siri [2], Y. Manasi [3], Y. Lekhya Naga Sai [4], G. Sai Swetha [5], P. Amrutha [6], Ch. Mouni Srinija [7]

Computer Science and Engineering, Sri Vasavi Engineering College, Tadepalligudem

santhirupa.cse@srivasaviengg.ac.in , jyothi6297483@gmail.com, manasi9346@gmail.com, lekhya38@gmail.com, saiswethagudala@gmail.com, amruthapachipulusu2k3@gmail.com, srinijachoppala@gmail.com

Abstract—Traffic flow prediction is an important aspect in an intelligent transportation systems. This model provides estimate the flow of traffic in a specific area at a specific point in an accurate manner. Studying traffic forecasting can help alleviate congestion and make travel safer and more cost-effective. Traditional models use flat networks, but recently, the vehicle count has increased rapidly. The main aim of this project is to predict how traffic will evolve overtime and allowing for better traffic management. Some of the Machine learning algorithms are Decision Tree, Random Forest, LGBM, e.t.c... The ab ove ML techniques use higher-level features from the raw input. We will discuss some of ML algorithms developed to address this problem. Because transportation systems are complicated, this information explains how different things affect these models and shows how well they work in different situations.

Keywords: Transportation systems, Traffic flow prediction, ML, Decision tree, Random forest, LGBM.

I. INTRODUCTION

The necessity for effective traffic prediction grows as cities expand and transportation needs increase. In this setting, traffic flow prediction becomes an important tool for streamlining traffic management plans, boosting transportation systems, and increasing commuter experiences. The main of this project is to develop a robust traffic flow prediction system that improves the performance of ML algorithms. By analyzing historical traffic data, environmental factors, and real-time sensor information, the project aims to create models capable of foreseeing traffic patterns with a level of accuracy and adaptability previously unattainable.

II. LITERATURE SURVEY

There are several models for predicting traffic flow such as XGBoost, SVM, Decision Tree.

There are a number of issues with Ganglong Duan's study on "Short Term Traffic Flow Prediction based on Rough Sets and Support Vector Machines," including low accuracy, high computing complexity, and challenges managing continuous data.

There are a number of difficulties with the research of Short-Term Traffic Flow Prediction Based on XGBoost in Xuchen Dong[3], including overfitting, handling imbalance data, hyper parameter tuning.

The investigation of the Gradient Boosting Decision Tree-based Freeway Travel Time Prediction Model has Data dependence, computing resources, especially with large datasets or when employing a huge no of trees in the ensemble, and difficulty when working with large datasets are some of Juan Cheng's constraints [5].

.

III. EXISTING SYSTEM

In existing system, authors proposed encompassing variables like traffic volume, vehicle speed, and temporal considerations by using several ML techniques like regression and ensemble learning for prediction. The existing model does not perform traffic prediction properly and also it requires a huge amount of time and money. So, in our project, We proposed a robust system that predicts vehicle congestion and time-based traffic flow to enable proactive measures for traffic management and optimization.

IV. PROPOSED SYSTEM

In respect to the persistent challenge of urban traffic congestion, we proposed a robust traffic prediction system that aims to predict and classify traffic conditions for effective optimization strategies. The system integrates advanced forecasting techniques with classification algorithms to anticipate vehicle congestion and time-based traffic flow accurately. By leveraging historical data and relevant infrastructure details, our system seeks to provide actionable insights for traffic authorities to implement preemptive measures and enhance overall traffic efficiency.

The proposed system utilizes a combination of historical traffic data and real-time sensor readings to develop precise forecasts of future traffic conditions. By using ML tecniques, particularly time series analysis, the system used to predict the temporal variations and traffic patterns. Additionally, employing classification models trained on labeled data enables the system to classify traffic conditions into distinct categories: high, heavy, low, or normal, based on historical trends and infrastructure characteristics.

By continuously updating its models with new data and feedback from implemented interventions, the proactive traffic management system ensures adaptability and reliability in predicting and classifying traffic conditions. This approach empowers traffic authorities with timely insights to implement preemptive measures, such as adjusting signal timings or deploying traffic diversion strategies, thus contributing to the alleviation of congestion and enhancement of overall traffic flow efficiency in urban environments.

V. METHODOLOGY

A. Dataset: The aim of this step is to collect the dataset from various sources like Kaggle.

B. Data Preprocessing: The collected data was then preprocessed to remove noise, missing values, and outliers. Feature engineering techniques were also applied to extract admissible information through the data.

C. Data Splitting: During training partitioning dataset is an essential step. It entails splitting the dataset into three subsets: validation, testing, and training. Here are few trivial ways for dividing data:

- *Train-Test Split:* This method evaluates a model's performance on a fresh dataset. Usually, between 70 and 80 percent of the comprehensive is allocated to training and 20–30% for checking out; Nevertheless, based on the use case and dataset this can differ.

- *Train-Validation-Test Split*: This approach assesses a ML techniques ability to generalize on fresh, untested data, in contrast to Train-Test Split. The version is trained using the training set, its overall performance is validated during training and hyper parameters are adjusted using the validation set, and its final performance is assessed using the testing set.

D. Model Training: The model underwent training utilizing the training set after the best-performing algorithm was chosen.

E. Comparing Model Accuracies: Using the right measures to grade the efficacy of models' performance is necessary for comparing their accuracies. The confusion matrix is a prevalent statistic employed for this purpose. Following the Decision Tree, Random Forest, XG Boost, LGBM, and SVR models' training and validation. where Random Forest and LGBM was high.

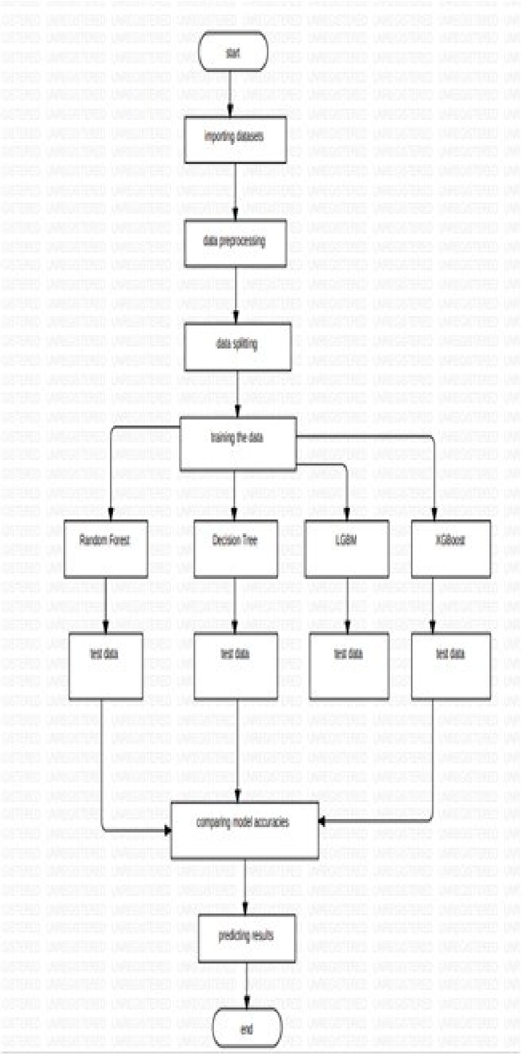

Fig.1 Methodology

VI. SYSTEM ARCHITECTURE

A. *Input Data (Traffic Flow Prediction Dataset):* This refers to the raw data with which the model is to be trained for predicting traffic flow is made to reference here. It probably contains historical traffic data together with weather, accident, and special event information.

B. *Data Preprocessing:* The current state of the raw data renders it unusable. To guarantee the data's integrity and consistency, this phase entails cleaning and formatting it. Erroneous or inconsistent data points can be fixed or eliminated.

C. *Feature Extraction:* All suitable attributes that will be useful for prediction are extracted from the preprocessed data. These variables could be the volume of traffic, the typical extracted from the preprocessed data. These variables could be the amount of traffic, the typical speed, etc.

D. *Scaling Features:* Scaling the characteristics to a common range is a chunk of this task. Similar-scale data features tend to yield better results.

E. *Splitting the data:* Datasets are separated out as Training along with Testing.

F. *Applying Machine Learning Classifiers:* The model is trained with the provided training data. Correlations with trends between the attributes and the tangible traffic conditions are recognised by the model.

G. *Visualization:* Utilizing this technique allows you to know about how the model's perform in comparison to the actual traffic circumstances.

H. *Comparing Model Accuracies:* In this stage we compare multiple models in order to obtain the one with the finest accuracy.

I. *Taking Model with Higher Accuracy:* The model that performs the best on the testing set is chosen to be the final traffic flow forecast model.

J. *Predicting Output:* The selected model is engaged to forecast the corresponding traffic circumstances whenever new, unseen traffic data becomes available. The result could be a more precise estimate of traffic volume or travel time, or it could be a classification (high, heavy, low, or normal traffic).

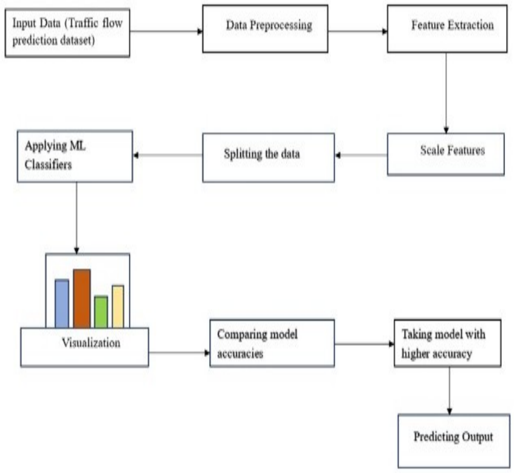

Fig.2 System Architecture

VII. SYSTEM OVERVIEW

A. *Data Acquisition:*

- Real-time traffic data from sensors (e.g., loop detectors, cameras).
- Historical traffic data.
- Public transport schedules.
- Weather data.
- Ride-hailing service usage data (optional).

B. *Data Processing & Analytics Engine:*

- Cleaning of data and preprocessing.
- Feature extraction (identifying relevant factors from raw data).
- Various models in machine learning : Traffic flow forecasting (predicting future congestion patterns).
- Traffic condition classification (categorizing traffic as high, heavy, low, or normal, with potential for more nuanced classifications).

C. *System Outputs:*

- Real-time traffic congestion forecasts for various time horizons.
- Dynamic traffic condition classification with visualization on a map.
- Alerts for high-risk areas with potential congestion.

D. *Integration & Applications:*

- Real-time signal optimization: adjusting traffic lights based on congestion forecasts.
- Integration with navigation apps: providing drivers with real-time traffic information for route planning.
- Variable message signs: displaying dynamic messages about upcoming congestion.

E. *Benefits:*

- Reduced travel times and improved fuel efficiency.
- reduced traffic before it starts.
- Enhanced safety through better traffic flow prediction.
- Informed decision-making for drivers and traffic authorities.

VIII. RESULT

The bar chart displays the results of six different predictive models and their accuracies: SVL, SVK, Random Forest, XGB, LGBM, and Decision Tree . The SVL model has the lowest accuracy of approximately 84.7 .The SVK model shows a slight improvement with an accuracy of 91.3. Both the SVK and SVL models have relatively low accuracy when compared to other four models. Whereas , the Random Forest, XGB, LGBM, and Decision Tree models exhibit significantly higher accuracy. Random Forest and LGBM models have the same accuracy value of approximately 98.3. The XGB model follows closely with an accuracy of around 98.1 and Decision Tree model has an approximate accuracy value of 97.2. Hence Random Forest and LGBM produces the highest accuracy value among all the models. In summary, the Random Forest and LGBM model demonstrates the highest accuracy, nearly related to Decision tree and XGB models. The SVL and SVK models have comparatively lower accuracy. These results suggest that Random Forest model and LGBM model are the most suitable choice.

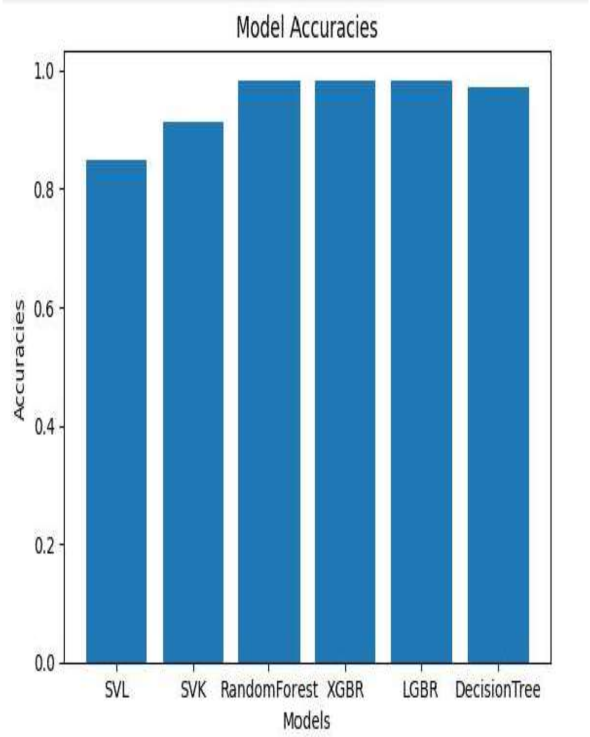

Fig.1 Comparing Model Accuracies

IX. CONCLUSION

Upon evaluating different machine learning algorithms, we observed that the random forest and LGBM models are the most accurate models for traffic prediction. Following this, we generated a confusion matrix, indicating an inclination towards random forest the substantial potential benefits include enhanced transportation systems and minimized economic losses.

X. REFERENCES

[1] Duan, Ganglong, Peng Liu, Peng Chen, Qiao Jiang, and Ni Li. "Short Term Traffic Flow Prediction Based on Rough Set and Support Vector Machine." In *2011 eighth international conference on fuzzy systems and knowledge discovery (FSKD)*, vol. 3, pp. 1526-1530. IEEE, 2011.

[2] Moses, Andrew, and R. Parvathi. "Vehicular traffic analysis and prediction using machine learning algorithms." In *2020 International Conference on emerging trends in information technology and engineering (ic-ETITE)*, pp. 1-4. IEEE, 2020.

[3] Hou, Yi, Praveen Edara, and Yohan Chang. "Road network state estimation using random forest ensemble learning." In *2017 IEEE 20th international conference on intelligent transportation systems (ITSC)*, pp. 1-6. IEEE, 2017.

[4] Chen, Yuanyuan, Hongyu Chen, Peijun Ye, Yisheng Lv, and Fei-Yue Wang. "Acting as a decision maker: Traffic-condition-aware ensemble learning for traffic flow prediction." *IEEE Transactions on Intelligent Transportation Systems* 23, no. 4 (2020): 3190-3200.

[5] Cheng, Juan, Gen Li, and Xianhua Chen. "Research On Travel Time Prediction Model of Freeway Based on Gradient Boosting Decision Tree." *IEEE access* 7 (2018): 7466-7480.